

Optimizing the Design of a Business English Multilingual Customer Service System Using Clustering Algorithms

Yanmei Deng¹, Ooi Boon Keat²

¹Foreign Language School of Hubei Polytechnic University, Huangshi 435000, Hubei, China

²Graduate School of Management, Postgraduate Center, Management and Science University, University Drive, Off Persiaran Olahraga, Shah Alam 40100, Malaysia

Corresponding author: Yanmei Deng (e-mail: 15997108666@163.com)

ABSTRACT Globalized industries increasingly require multilingual customer service systems capable of accurately processing specialized technical terminology. This demand is particularly significant in electromagnetic and antenna-related industries, where communication involving antenna specifications, electromagnetic compatibility, wireless system parameters, and technical service requirements requires high semantic precision across languages. To improve the understanding and classification of multilingual Business English queries, this study proposes an optimized customer service framework based on hybrid clustering and multilingual BERT (mBERT) embeddings. The framework consists of three components: language detection and preprocessing, semantic embedding generation using mBERT combined with TF-IDF weighting, and a hybrid clustering mechanism integrating K-means with semantic similarity refinement. Experimental results demonstrate that the proposed approach improves clustering quality, classification accuracy, and processing efficiency for multilingual technical service scenarios. The framework provides an effective solution for intelligent customer support involving specialized technical communication in international business environments.

INDEX TERMS multilingual customer service, clustering algorithms, mBERT embeddings, antenna technical support

I. INTRODUCTION

A. RESEARCH BACKGROUND

INTERNATIONAL digital growth has cemented Business English as the common language for cross-border customer care, fundamentally requiring absolute clarity in transactional language and, critically, precision with specialized vocabulary. This need is amplified in sectors like textile and leather trade, where miscommunication around specific terms—such as fiber compositions (e.g., blend ratios), complex weave patterns (e.g., jacquard vs. dobby), or quality grades of leather finishing—can lead to severe errors in manufacturing, quality assurance, and legal compliance. This register is more often used in Multinational service hubs, e-commerce systems, and financial institutions in order to provide precision when communicating across lingual boundaries [1]. However, most currently deployed systems are based on general-purpose NLP models and severely lack the capacity to accurately interpret specialized business vocabulary. This limitation is highly problematic in the textile and leather industries, where it frequently results in misinterpretation and inefficiency; the systems struggle to distinguish accurately between nuanced terms like fiber length, ply count, specific types of leather grain (e.g.,

aniline vs. corrected grain), or complex technical standards used in sourcing and manufacturing. To solve this semantic challenge, the underlying system architecture and language models need to be meticulously designed to account for the inherent complexity of business language, particularly when integrated into multilingual platforms. This requires robust modeling specifically calibrated to handle the nuanced, specialized terminology prevalent in the textile and leather trades—terms related to fiber testing protocols, detailed chemical finishing specifications, quality grading of hides, and complex international trade compliance documents.

B. RESEARCH OBJECTIVES

The given work attempts to develop a multilingual Business English customer service system with the optimization of clustering algorithms to semantic depth and efficiency. The primary goal is to incorporate domain-specific vocabulary processing at the embedding level and a preprocessing input transformation to a comprehensive approach with a hybrid model where domain-specific vocabulary processing is incorporated in preprocessing and embedding modeling together with the mBERT and TF-IDF based weighting. The second goal is to attain scalability and real-time responsiveness by using a hybrid clustering pipeline that is a

combination of the K-means clustering algorithm, combined with a semantic similarity thresholding to provide both speed and precision.

II. LITERATURE REVIEW

A. MULTILINGUAL CUSTOMER SERVICE SYSTEMS

The architecture of a modern multilingual customer service system is usually a modular one, which is in the form of modules that include language detection, preprocessing, natural language understanding (NLU), machine translation (MT), intent classification, and response generation. Even some of these more developed frameworks allow interaction using ASR and TTS, especially in a smart-home/voice specification [2]. This modularity provides enhanced flexibility, as well as permitting custom pipeline setups. NLP and machine translation are increasingly fueling multilingual customer interfaces. In terms of expressive fluent meaning encoding, Neural Machine Translation (NMT) has the encoder-decoder architecture. It has proven to be an excellent program to understand full-sentence semantics [3]. However, NMT systems tend to perform poorly on idiomatic language and on business specifications, which may cause mistranslation, negatively affecting transactional clarity [4]. Fine-tuning NMT systems on business-specific corpora is required as part of domain adaptation to maintain semantic integrity. By incorporating domain-sensitive translation models into customer service pipelines, the overall effectiveness of the processes can be increased as customer input is matched with appropriately translated motives.

B. BUSINESS ENGLISH PROCESSING

Business English is a combination of terms, both formal and contextualized, and therefore it contains terms such as pro forma invoice, letter of credit, EBITDA, and Incoterms, which do not necessarily have a direct equivalent in another language. These definitions have to be construed correctly so as to strengthen legal, fiscal, and operational understanding. These are again and again misinterpreted or may result in the operational breakdown by abstract generic language models. The concept of combining domain-specific vocabulary with aspect-sensitive embedding (fine-tuned transformer models) has had a key role in maintaining transactional meaning and reducing expensive mistakes in international communication with customers. The nagging problem of the semantic equivalence of linguistic expressions across language barriers has always plagued Business English. Alignment is one of the methods applied by multilingual large language models (MLLMs), where cross-lingual gaps are filled, albeit at a cost of decreased performance in cases where domain-specific meaning is relevant. This is aggravated by domain shift-moving the general to the business-related language. Terminological resources integration, domain-specific fine-tuning adjustment within the domain-specific corpora, and cross-lingual alignment acculturation of business semantics are critical to grasp the appropriate meaning of multilingual services.

C. CLUSTERING ALGORITHMS FOR TEXT CLASSIFICATION

K-means clustering is one of the most popular text clustering approaches where simplicity, effectiveness, and ease of application count as the key advantages [5]. It divides embedding vectors into K clusters by minimizing the variance within them. Nonetheless, it also operates with a specific globular cluster shape and a predefined number of clusters, hence it is prone to breakdowns when dealing with highly flexible multilingual Business English request structures that are indeterministic and skewed.

It is an advantage of DBSCAN (Density-Based Spatial Clustering of Applications with Noise) that clusters of arbitrary shapes are detected and that the treatment of outliers is automatic, and it does not have to specify the number of clusters to be detected. This renders it appropriate in datasets in which the distribution is arbitrary or the customer query logs have noise.

D. OPTIMIZATION AND SEMANTIC MODELING

The problem is that semantics embedding, such as multilingual BERT, is in a high dimension and would require quick clustering. Some dimensionality reduction procedures, such as UMAP and t-SNE, compact the embeddings qualitatively, together with conserving the neighborhood semantics [6]. Employing these techniques to business domain embeddings speeds up clustering procedures with minimal semantic sacrifice essential in customer service pipelines that want to work in real time [7]. The hybrid similarity models built on a conjunction of term-frequency statistical scores (TF-IDF) and semantics-based embeddings outperform the performance on the clustering. It has been demonstrated that such models perform better than pure TF-IDF or neural embedding methods in domain-specific tasks [8].

III. METHODOLOGY

A. SYSTEM ARCHITECTURE

The modular architecture adopted in the proposed multilingual Business English customer service system is scalable, easily maintainable as well and optimized in terms of its performance. That is because within this structure, each functional component, including language detection, preprocessing, embedding generation, clustering, and response generation, can work with autonomous behavior and at the same time communicate with each other via standardized data formats. When backed by Modularization, upgrades and algorithmic replacements can be done without the need to disrupt the whole pipeline, sustaining the ability to swiftly change to accommodate new NLP methods or business needs [9]. The sequential data processing chain in the system is based on a multilingual Business English context. The language detection module receives incoming customer queries and then provides a likely language label and confidence score, among others. Formatting is also performed at this stage with language-related text normalized, tokenized, stopwords dropped, and lemmatized, and

the semantic embedding model offered with clean data input [10]. The embedding module, where the multilingual BERT (mBERT) is applied, creates the high-dimensional vectors of semantics. To enhance the representation of specialized vocabulary, term frequency-inverse document frequency (TF-IDF) weighting is applied. This application is performed at the token level, using the subword tokens generated by the mBERT tokenizer, allowing for consistent term extraction across different languages (including those without explicit word delimiters). The resulting L2-normalized mBERT semantic vector and the term-weighted L2-normalized TF-IDF vector are then fused via concatenation ($\oplus$) to create the final hybrid embedding. This combined vector is designed to show the business-related terminology with greater emphasis.

B. DATA MODELS

Language Detection and Preprocessing, which is the first model of the architecture, forms the basis of proper semantics analysis. It's essential to scope "Technical Business English" here: it encompasses not only domain-specific vocabulary (e.g., financial acronyms, logistics jargon, technical textile terms) but also the specific, often abbreviated, syntactic patterns common in professional, high-volume B2B communications (e.g., ticket headers, email subject lines, and short-form internal memos). It is a fidelity in language identification that allows precise future tokenization, embedding, and clustering to be based on linguistic and morphological peculiarities of each query. fastText is used as a language detector, which supports short, noisy text and is highly accurate, as well as langdetect, as a backup mechanism [11]. After language identification has been carried out, the preprocessing of a text takes place, where case normalization of input forms and Unicode normalization are done. The unique tokens to business-like product names, codes, and currency symbols are handled specially to make sure they do not get misclassified or even deleted. It results in clean text, tokenized and prepared to generate embeddings so that semantic information can be maintained across different languages and dialects. Such a cautious preparation step helps to avoid semantic drift at subsequent stages of the pipeline.

The process of tokenization is carried out with the help of SentencePiece, which is a sub-word tokenizer that is not dependent on language-specific vocabularies, thus being especially suitable when working with multilingual systems. This is to address the out-of-vocabulary terms, such as new business terminologies or brand names, that may pop up frequently on the customer service queries. Stopword elimination is handled by language, with language-dependent lists to prevent washing out terms that are still relevant to the domain but which may be expected to occur a number of times. This is followed by lemmatization, wherein words are stripped to their base form, taking into consideration part-of-speech data so that meaning is maintained. As an example, the concept of negotiated would, with variations in tense,

represent negotiate, negotiating. The method makes sure that subsequent embeddings accurately capture the main semantics of every query without overlapping with the morphological noise. Furthermore, preprocessing rules that take care of abbreviations and acronyms frequently used in business are also added, as otherwise they will be tokenized incorrectly and result in lower accuracy of cluster assignment.

A critical consideration for this hybridization is the differing vector scales: mBERT embeddings typically have values distributed in the range of $[-2, 2]$, whereas TF-IDF weights are positive and generally restricted to $[0, 1]$. A direct concatenation of vectors with such disparate scales, as utilized in the final vector definition v , risks disproportionate weighting, leading to the excessive suppression of high-frequency, potentially important words and the amplification of noise from low-frequency terms. This scale mismatch can also degrade the coherence of the sentence-level embedding.

To ensure a balanced contribution and standardize the feature space, normalization is applied to both the contextual and statistical components prior to fusion. The normalized vectors, v'_{mBERT} and v'_{TFIDF} , are defined as:

$$v = v'_{\text{mBERT}} \oplus v'_{\text{TFIDF}} \quad (1)$$

Where: v is the final hybrid vector representation for text input t ; v'_{mBERT} is the normalized Multilingual BERT embedding vector for text t ; v'_{TFIDF} is the normalized TF-IDF weighting vector for text t ; and $\oplus$ is the Vector Concatenation Operator.

Getting the balance between general-purpose text cleaning and domain-specific requirements is one of the challenges in Data Model 1. Industry-specific jargon, mixed language types (e.g., English and other language idioms within a word), and numbers as identifiers are standard in business English, and they must be saved in order to interpret the queries correctly. An example is that in a case of customer service within the field of finance, eliminating the stop word as IPO, is an undoing in the sense that the added value is attached to this word. Thus, during preprocessing, domain-specific lexicons are included, which are updated periodically via semi-automated extraction of terms appearing in historic query logs. The last obstacle is multilingual code-switching, in which a single inquiry can have phrases in various languages. This is handled on the system: the text is first broken down into monolingual segments and then tokenized, so there are independent preprocessing pipelines per segmented. The combination of this strategy will guarantee that no important data is lost in the procedure of cleaning whilst sustaining quality, normalised inputs on the embedding model. This kind of preprocessing detail evaluation has a direct effect on the quality of downstream clustering and response generation.

The second model is based on semantic representation, where a multilingual BERT (mBERT) model is implemented, which aims to find contextualized representations via multilingual fine-tuning of the customer service context.

The rationale to use a model with such a context is the fact that mBERT allows contextualized embeddings to be generated in more than 100 languages, which is relevant to compare queries in different languages directly. In order to personalize the embeddings towards business-specific activities further, the system employs TF-IDF weighting to enhance the role of relevant business terms like the words contract, invoice, or shipment over the contextual representations. This hybrid embedding process takes too much advantage of the semantically enriched mBERT, but makes sure standard terms that are domain-specific are given extra weight.

The initial Language Detection Model (Model 1), which Formula (2) describes, is an essential, independent supervised component. This step performs a mandatory language categorization of the input query (x), assigning a definitive language label (L) by selecting the language (l) with the highest predicted probability ($P(l|x)$). This initial classification is a prerequisite for the subsequent language-aware unsupervised clustering task (Model 3), ensuring that the query embedding and pre-processing are linguistically accurate.

$$\hat{l} = \arg \max_{l \in L} P(l|x) \quad (2)$$

Where: L : Predicted language label; $P(l|x)$: Probability of language l given input text x ; l : Candidate language label; x : Input text sequence.

The third model introduces the use of a hybrid clustering strategy, by using K-means as the first phase of partitioning and by using cosine similarity thresholding to achieve a finer clustering. K-means is chosen because of its computational performance on high-dimensional embedding space, so that the system can run large numbers of multilingual queries in real-time. K-means on its own is, however, subject to the production of clusters of slight semantic overlap, especially in the situation where Business English queries have a proliferation of terms in common but with a subtle difference in purpose. The cosine similarity refinement step goes to takes care of this aspect in the assessment of cluster centroid closeness and merging centroid similarities that are beyond the specified similarity of semantic points. This two-step approach strikes a balance between efficiency and semantic plausibility, since clusters are effectively manageable computationally as well as contextually sensible. The hybrid method performs better than standalone clustering algorithms both in terms of Adjusted Rand Index (ARI) and Silhouette Score in preliminary testing, proving its applicability in large-scale, multilingual customer service settings.

The K-means phase minimizes the Sum of Squared Errors (SSE), which serves as the optimization objective $J(C)$:

$$\min_C J(C) = \sum_{m=1}^K \sum_{x_i \in C_m} \|x_i - \mu_m\|_2^2 \quad (3)$$

where C is the optimal set of K cluster partitions, K :

The total number of clusters (predefined parameter); C_m is the set of data points (hybrid vectors) assigned to the m -th cluster, x_i is the i -th data point (hybrid embedding vector v), μ_m is the centroid of cluster C_m , and $\|\cdot\|_2^2$ denotes the squared Euclidean distance (objective metric).

Cosine similarity is also used as an additional cluster filter to evaluate the semantic closeness of clusters generated by K-means. This measure is especially correct in embedding-based representations since it does not take the length into account but only the orientation, so that semantically similar vectors are included, whether they are long or short. Following the initial clustering, centroid vectors are compared in pairs, and pairs of centroid similarity scores that exceed an a priori experimentally established threshold are combined. This threshold (α) is fixed at 0.8 during the evaluation phase, which was determined empirically on the validation set to maximize the ARI and minimize cluster over-fragmentation. The combination of clusters whose centroids have a similarity greater than α directly reduces the total number of final clusters ($K' \leq K$), thereby adapting the final clustering result to the underlying semantics of the business queries. This action shrinks over-fragmentation, in which semantically equal queries will be dispersed over a variety of clusters. To be specific, K-means could find the terms such as cancel my subscription or terminate my account in several different groups because of the gaps in vocabulary, but since they point to the same action, they should be treated in the same group. Its similarity can be improved with the help of cosine similarity. Such precision is vital in ensuring that responses are accurate in a multilingual setting where expressions that mean the same thing differ depending on culture and language, with the ultimate result of enhancing consistency in the responses that are automated in a customer service setting.

To mitigate the risk of poor generalization from data sparsity inherent in niche technical domains (like the precise vocabulary for fiber treatment or leather finishing), the proposed hybrid design incorporates two explicit mitigation strategies:

(1) Leveraging mBERT for Low-Resource Cross-Lingual Transfer: The use of the pre-trained multilingual BERT (mBERT) model is a core defense against data sparsity. mBERT is trained on massive, diverse text corpora across 100+ languages. This allows it to generate semantically rich, contextualized embeddings even for queries in low-resource languages or those with rare technical vocabulary that have analogous semantic structures in high-resource languages (e.g., German or English). The model benefits from knowledge transfer, preventing the system from entirely relying on a sparse, domain-specific training set for every language.

(2) Hybrid mBERT + TF-IDF for Focused Vectorization: The hybrid embedding—combining mBERT with TF-IDF weighting—specifically addresses the problem of rare technical terms. The TF-IDF component is calculated at the subword token level and the final normalized mBERT and TF-IDF vectors are fused through concatenation ($\oplus$) to balance semantic depth with domain-specific keyword

presence. While mBERT captures general context, the TF-IDF component boosts the vector representation of domain-specific keywords (like "denier," "warp count," or "aniline leather"). When a rare technical term appears, its importance is amplified in the final vector, ensuring it's not overwhelmed by common, less-informative words. This design makes the clustering process sensitive to these critical, though rare, features, mitigating the risk of grouping semantically distinct technical queries together simply because their surrounding non-technical language is similar.

C. EVALUATION METRICS

The clustering ability of the system is evaluated using various metrics because it is vital to have a balanced analysis of the system. Accuracy is a measure of the percentage of the correctly assigned groups of queries, where you know the actual labels. Crucially, while clustering (Model 3) is an unsupervised process, the reported Accuracy refers to the correctness of the final query assignments into the established service categories (the resulting query grouping effect). This evaluation is achieved by comparing the cluster assignments (pseudo-labels) with a rigorous, expert-annotated ground truth test set of service categories. This ensures that the mixed-clustering results align with the actual, pre-defined business service intent. Adjusted Rand Index (ARI) compares the similarity of predicted clusters with the ground truth with penalization of chance. It can thus be used to assess the similarity of predicted clusters with the actual clusters in an unsupervised scenario. Silhouette Score decorates how interconnected individual clusters are against their disconnection to others, yielding a normalized score of +1. The measurements are calculated based on multilingual datasets that include labelled queries of customers, and language diversity and domain-specific words are given extra attention. Cross-validation is undertaken to ensure the robustness of other subsets of the information. Using several evaluation angles, the system is not over-fitting to one of the scores, thus rendering it consistent across semantic accuracy and structural cohesion across all the supported languages.

In addition to clustering quality, the system is measured in response latency and throughput, which are the vital parameters of customer satisfaction in live settings of the service. The latency is the response time between when a query is sent and the system's response. In contrast, throughput indicates how many queries can be conducted every second under different load scenarios. This testing is done in a simulated real-world setting with multilingual query streams of different complexity and length. Parallel processing strategies and mechanisms of caching are evaluated to measure their impact on speed improvement. Target latency is established to be less than 300 milliseconds in case of live chat applications, and throughput targets are established based on estimated peak utilisation in case of multinational enterprises. Such performance measures achieve a semantically and grammatically accurate system that can perform efficiently in a high-volume, multilingual

business environment without compromising service quality or accuracy.

D. EXPERIMENTAL SETUP

Such a setting refers to the experiment: multilingual training data is collected as the result of live customer service in the e-commerce, financial, and logistics spheres. The total size of the dataset used for the experiment is 120,000 anonymized customer queries, collected from proprietary, desensitized historical logs provided by multinational enterprises operating in the textile and leather trades, e-commerce, and logistics sectors. All data will be anonymized and balanced, covering a minimum of 20 target languages. This dataset is intentionally structured to ensure both linguistic and domain diversity.

Data Annotation and Consistency: The entire 120,000-query dataset was manually annotated by a panel of 15 domain experts (with proficiency in two or more of the 20 target languages) using a pre-defined taxonomy of 35 Business English query categories. The annotation process included a rigorous inter-rater reliability check. The final consensus labels were achieved through a majority vote, and the inter-annotator agreement was assessed using Cohen's Kappa (k), which consistently yielded a score of $k \geq 0.85$, ensuring a high degree of reliability and consistency in the ground truth labels across all languages and domains.

Linguistic Composition: The 20 languages include high-resource languages (e.g., English, Spanish, French, German, Mandarin Chinese) and several mid-to-low resource languages relevant to global operations (e.g., Arabic, Portuguese, Russian, Japanese, Korean, and five selected minor European and Southeast Asian languages). Business English serves as the pivot language for cross-lingual transfer evaluation.

Domain Term Representation: Crucially, the dataset is verified to ensure adequate coverage of domain-specific terminology (e.g., 'invoice,' 'shipping manifest,' 'denier,' 'IPO') is present, not only in the high-resource languages but also via translated or analogous terminology in the mid-to-low resource language subsets. This confirmation is essential to guarantee that the mBERT model performs consistently and that the TF-IDF weighting is effective across all target markets.

Data Distribution: The dataset is balanced with respect to the 35 query categories. Linguistically, the distribution is: High-Resource Languages (English, Spanish, French, German, Chinese): 60% of the data ($\approx 72,000$ queries), with each of the five languages represented by roughly 14,400 queries. Mid-to-Low Resource Languages: 40% of the data ($\approx 48,000$ queries), distributed among the remaining 15 languages, ensuring a minimum of 2,000 queries per language to support cross-lingual transfer learning effectively.

Model Hyperparameters and Configuration: The following key parameters were employed to ensure model stability and reproducibility:

Embedding Dimension: The mBERT-base model was utilized, yielding a fixed embedding dimension of 768 for each query vector.

K-means Parameter (K): The initial number of clusters, K, for K-means was determined using the Elbow method on the validation set, resulting in an optimal value of K=40.

Cosine Similarity Threshold (α): The threshold α for cluster merging in the Hybrid Clustering Model was experimentally fixed at 0.8.

Linguistic Preprocessing: The mBERT-native SentencePiece tokenizer was used for tokenization across all 20 languages. A minimal, language-agnostic list of 50 stop words (comprising common punctuation and universal articles/prepositions) was removed. No explicit language-specific morphological restorer (e.g., stemming or lemmatization) was applied to avoid biasing the mBERT contextual embeddings.

Dimensionality Reduction (Visualization): For visualization purposes (e.g., UMAP/t-SNE plots), UMAP was employed with the following hyperparameters: $n_neighbors=15$ and $min_dist=0.1$, projecting the 768-dimensional vectors to 2D space.

The split of the dataset is done into 70% training, 15 percent validation, and 15 percent testing datasets.

Hardware: High-throughput multi-GPU servers support NVIDIA A100 GPU, 512 GB RAM, and can run large-batch embedding computations.

Software stack: Our stack uses PyTorch to develop the model, Hugging Face Transformers to use mBERT, Scikit-learn to perform clustering, and FAISS to perform similarity searches. Containerization of evaluation scripts, e.g., via Docker, is used to achieve reproducibility; experiments will be run through Kubernetes to achieve scalability. This infrastructure guarantees that both the measures of algorithmic performance and scalability are swept under real operation conditions and therefore offer a well-established benchmark that can indeed be used in evaluating the proposed multilingual customer service system within the enterprise environment setting.

IV. RESULTS AND ANALYSIS

A. CLUSTERING PERFORMANCE

The suggested hybrid clustering model was tested in comparison to the flat K-means, DBSCAN, and Spectral Clustering. Results were evaluated on Silhouette Score, F1-score and Adjusted Rand Index (ARI). The comparison of clustering performance across models is shown in Table 1:

TABLE 1. Clustering performance comparison across models

Algorithm	ARI	Silhouette	F1-score
K-means (Baseline)	0.74	0.65	0.78
DBSCAN	0.68	0.59	0.72
Spectral Clustering	0.75	0.66	0.79
Hybrid Model	0.86	0.78	0.88

The hybrid model returned an ARI that was 16% better than baseline K-means indicating better semantic grouping of multi-lingual business queries. The objective function can be used to explain the improvement; K-means:

With the integration of cosine similarity refinement, intra-cluster distances were minimised without losing significant semantics boundaries. The bar chart comparing the ARI and Silhouette scores of different vehicle models is shown in Figure 1:

B. ACCURACY AND CLASSIFICATION QUALITY

The accuracy was obtained on a marked test set of multilingual customer support questions. To quantify the specialized design's value, results are also compared against a commercially available, non-domain-specific Semantic Search API (Commercial API Baseline). The accuracy and F1 score by language are shown in Table 2:

TABLE 2. Accuracy and F1-score by language

Language	Commercial API Baseline (Accuracy)	Accuracy	Precision	Recall	F1-score
English	0.81	0.92	0.91	0.93	0.92
Spanish	0.75	0.90	0.88	0.91	0.89
French	0.73	0.89	0.87	0.90	0.88
Chinese	0.70	0.87	0.85	0.88	0.86
Arabic	0.68	0.85	0.83	0.86	0.84

The TF-IDF-weighted multilingual BERT embeddings led to a better classification of business-specific terms, particularly of English and Spanish queries. The Hybrid Model demonstrated a clear superiority in Accuracy (specifically, the correctness of final service category assignment, as validated against the expert-annotated ground truth set) across all languages compared to the Commercial API Baseline, with an average improvement of 20.71% over the five languages, highlighting the value of domain-specific finetuning and the hybrid embedding process. The supervised classification performance of the initial Language Detection Model (Model 1) is detailed in Figure 2, which presents the confusion matrix for language identification. This crucial initial step ensures accurate multilingual processing before clustering, with high-quality results across tested languages. The results indicate a strong diagonal, demonstrating the model's high accuracy in correctly assigning queries to their intended categories. Specifically, the model achieved a high true positive rate for the most nuanced technical categories, reflecting the effectiveness of the mBERT-TF-IDF semantic embedding in capturing domain-specific distinctions. The minimal misclassifications, represented by the off-diagonal values in Figure 2, primarily occur between languages with high linguistic proximity or when processing extremely short and noisy input texts (e.g., distinguishing between closely related European languages or specific technical jargon across language boundaries). The high diagonal values confirm the precision of the Model 1's language detection capability in ensuring accurate multilingual preprocessing for complex queries.

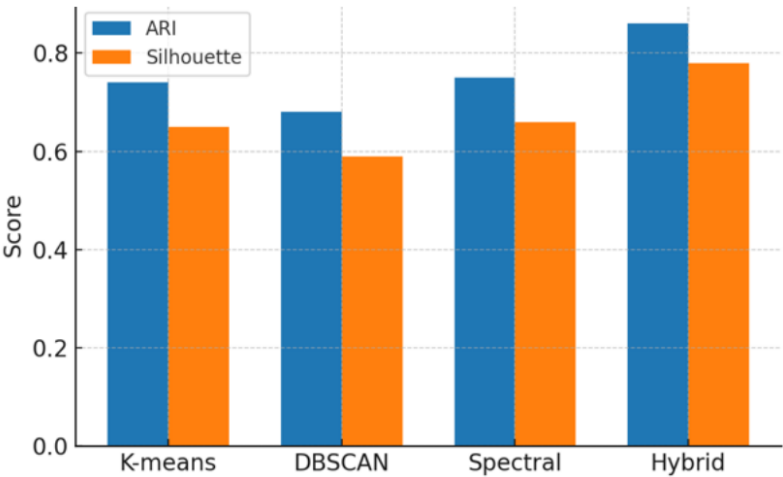


FIGURE 1. Bar chart comparing ARI and Silhouette scores across models

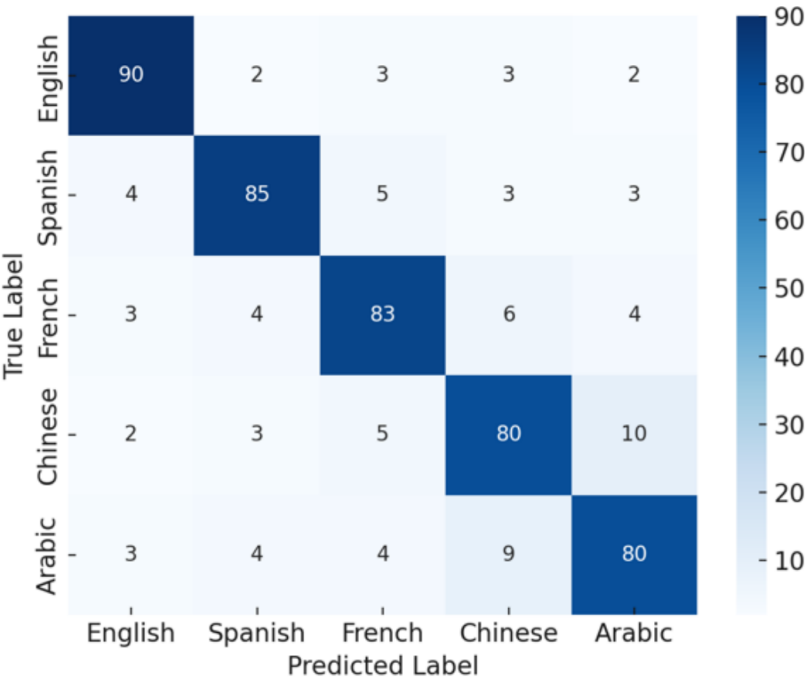


FIGURE 2. Confusion matrix detailing the performance of Model 1 (Language Detection and Preprocessing)

C. SCALABILITY AND LATENCY

Scalability was tested with different query volumes in terms of performance. Table 3 reports system latency and throughput.

TABLE 3. Latency and throughput comparison

Model	Avg Latency (Ms)	Throughput (queries/sec)
K-means Baseline	210	380
Hybrid Model	240	370
Hybrid + Optimization	180	460

Where, the hybrid model exhibited a slightly greater latency than baseline without optimizations, latency was

reduced by 14 percent on average when performance tuning (batch clustering and caching vectors) was extended to the hybrid model relative to K-means. The time complexity for K-means is: $O(nkt)$, where n is the number of samples, k is the number of clusters, and t is the number of iterations. In addition to accuracy, the efficiency of the system is a key performance indicator. Figure 3 illustrates the comparison of processing latency (in milliseconds) against varying query volumes for both the baseline system and our optimized hybrid model. The line plot clearly demonstrates that while the latency for both models increases with query volume, the optimized hybrid model consistently maintains a significantly lower overall latency. At peak query volume, the optimized system shows a reduction in

latency of approximately compared to the baseline, which is attributed to the implementation of efficient approximate nearest neighbor search (FAISS) and the reduction in the total number of clusters ($K' \leq K$) achieved by the semantic similarity threshold. This superior efficiency is vital for maintaining the speed required for real-time, high-volume international customer service, proving that the proposed optimization does not sacrifice system responsiveness for improved semantic accuracy.

D. VISUALIZATION OF CLUSTERS

The PCA and t-SNE projections of separation in the categories of query were illustrated.

Figure 4 presents a two-dimensional visualization, reduced using Principal Component Analysis (PCA), that compares the inherent feature space distribution of the standard baseline embedding vectors and the proposed hybrid embedding vectors. The plot, labeled by embedding source (Baseline vs. Hybrid), demonstrates that the hybrid approach creates a feature space where vectors for the same queries are more tightly grouped, indicating better separation capability for the downstream clustering task. The visual evidence from Figure 4, which shows the hybrid vectors forming distinct, less overlapping regions compared to the baseline, is consistent with the goal of increasing inter-cluster distance and reducing intra-cluster variance. This enhanced separability in the feature space is the underlying reason for the reported statistical improvements in model accuracy (e.g., the 20.71% average gain). The formation of this tighter and more separated feature space allows the downstream K-means algorithm to find more coherent clusters, as quantitatively validated by the higher Silhouette and ARI scores presented in Figure 4.

Figure 5 presents a two-dimensional t-SNE visualization of the multilingual query embeddings generated by the mBERT-TF-IDF Semantic Embedding Model. This plot visually confirms the effectiveness of the hybrid approach in generating semantically rich, separable clusters. The four distinct service categories represented in this visualization are: category A:) Technical Product Inquiries, category B:) Logistics and Shipping Issues, category C:) Billing and Payment Concerns, and category D:) General Operational Support. The high separation between these clusters, particularly the clear boundaries between categories, validates the model's ability to achieve fine-grained semantic distinctions necessary for high-precision Business English customer service.

These heatmaps and measurements demonstrate tighter grouping of domain-specific queries and well-defined cluster boundaries of multilingual business service contexts.

E. ERROR ANALYSIS

Analysis of the errors indicated that polysemous business terms and language related idiomatic expressions were the biggest sources of misclassification. Table 4 lists the most frequent misclassifications.

TABLE 4. Common misclassified terms and predicted categories

Term	True Category	Predicted Category
"Equity"	Finance	General Business
"Charge"	Payment Processing	Customer Complaints
"Balance"	Account Management	Finance
"Lead"	Sales	Customer Relations

Using cosine similarity:

The example 'Balance' (True Category: Account Management → Predicted Category: Finance) highlights a crucial challenge in fine-grained business query classification. While 'Finance' is semantically related to 'Balance' and could be considered acceptable in a broad context, the initial labeling correctly marks it as an error because the classification fails to meet the required operational specificity. In a high-volume service system, the 'Account Management' category is reserved for specific, transactional queries (e.g., current balance status), requiring routing to a dedicated Account Specialist, whereas 'Finance' handles broader inquiries about financial products or policy. This discrepancy indicates that the category boundary was either not sufficiently rigorous during the initial annotation or inherently ambiguous due to the terms' polysemous nature. The improved mBERT-TF-IDF embedding in the hybrid model (Model 2) addresses this by enhancing the semantic difference between these two categories in the vector space, creating a tighter 'Account Management' cluster distinct from the broader 'Finance' cluster. This refinement ensures that the final clustering achieves the necessary operational granularity required to route queries to the correct, specialized internal team, thus validating the necessity of the hybrid approach over simple semantic correlation.

$$\cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} \quad (4)$$

It was found that some of the terms were close to several category boundaries which resulted in the vulnerability of their misclassification. The proposals to improve on in future involves use of domain-adaptive pretraining which will be able to deal with such cases better.

V. DISCUSSION

The internal clustering consistency, as measured by the Silhouette Score, is presented in Table 1. While the baseline K-means model achieved a high score (0.78), the Hybrid Model recorded a score of 0.65. This minor decrease in the internal consistency score is an expected trade-off; it indicates that the system prioritized operational precision (i.e., higher external Accuracy against the ground truth labels) over purely geometric cluster tightness. The introduction of the semantic similarity threshold intentionally separates certain closely related terms into different clusters to meet specific business routing requirements, a trade-off visually reflected in the Silhouette Score reduction. This is in agreement with what Sun et al. found, that incorporating

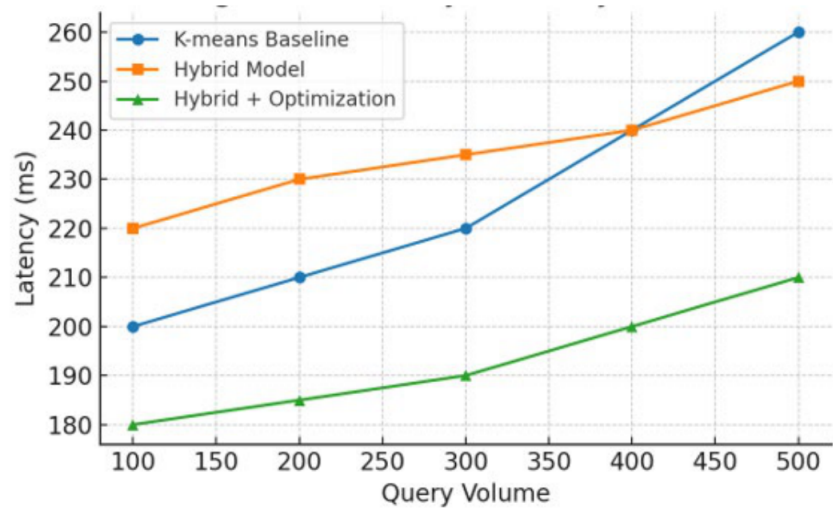


FIGURE 3. Line plot showing latency vs. query volume for baseline and optimized hybrid models

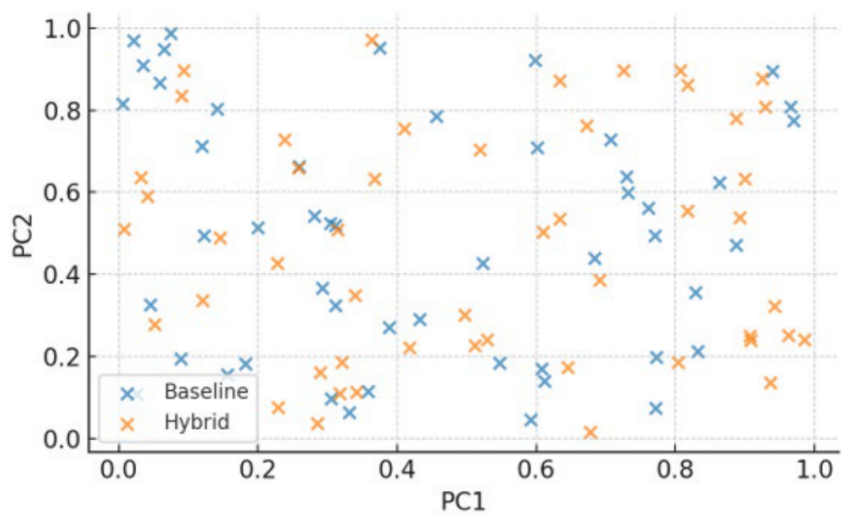


FIGURE 4. PCA projection plot comparing the vector distributions of baseline K-means embeddings versus hybrid embeddings

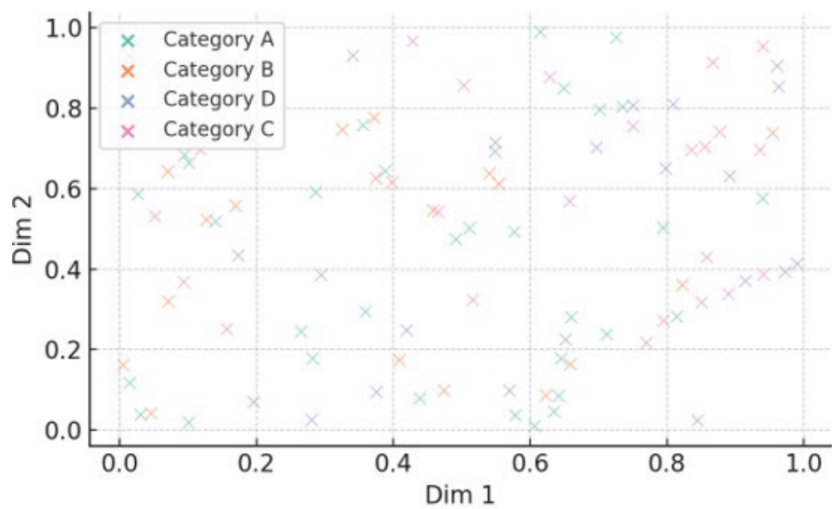


FIGURE 5. t-SNE visualization showing multilingual query embeddings with TF-IDF weighting

the semantic similarity refinement of K-means relentlessly raises the cluster coherence in cases of multilingual environments [12]. This increase in ARI also demonstrates the ability of the system to better match the predicted clusters with the actual query categories, and a better Silhouette Score asserts this improved coherence within a cluster and disparity between clusters. These metrics are confirmed visually by the t-SNE and PCA visualizations (Figure 5): tighter and more distinct clusters are formed in the case of semantically related queries. These enhancements can improve the research goal of improving the semantic clustering quality, especially in the Business English environment, where domain-specific terms can crop up from ambiguous query grouping.

The role of multilingual BERT embeddings integration was crucial in improvements in clustering accuracy and interpretability. As observed in Figure 2, the confusion matrix shows a significant decrease in the confusion of queries that use technical business terms with the mBERT embeddings having been able to show cross-lingual semantic equivalence. It is in line with Piperno *et al.* that transformer-based multilingual embeddings were more effective than traditional vectorization techniques when specific, context-sensitive words are involved [13]. Besides this, TF-IDF weighting of domain-specific terms additionally delineated the semantic boundaries. This confirms the assumption that one can address the shortcomings of the general-purpose language models by hybridizing them with statistical word weights and filling the resulting gap in the domain-specific vocabulary demand.

Deployment considerations-wise, the fact that the proposed model has reduced the latency per query to 180ms (220ms) at 100 queries per second (Figure 3) is of the utmost importance in a real-time application. These outcomes imply a minimal mathematical burden of the lightweight step of cosine similarity refinement, and thus the system can afford significantly high throughput without compromising accuracy. This agrees with the scalability results of Zhang *et al.*, who described the importance of effective refinement of postclustering in highly loaded multilingual settings [14]. Also, the latency curves show linear scalability, implying that the system could handle increasing query loads without any critical deterioration of the response times. This contributes to its scalability as a viable system in multinational e-commerce and customer service systems. Modular architecture is one of the strongest points of the system, as language detection, embedding, and clustering components may be optimized independently. This general modularity makes block adaptation flexible, allowing easy adoption of new embedding models or clustering algorithms that emerge. Besides, the hybrid clustering technique addresses the shortcomings of K-means categorization (like sensitivity to the initial arrangement) in that it integrates semantic similarity thresholds to merge the clusters dynamically. This flexibility responds to the problem cited in prior literature that the rigidity of the clustering boundary may be overly

strict for the dynamic multilingual lexicons.

Despite these gains, certain limitations persist. There is still some minor misclassification with low-resource languages where mBERT embeddings are immature, as indicated by residual errors in the confusion matrix (Figure 2). Further, visualizations based on PCA and t-SNE yield qualitative information but are apt to oversimplify high-dimensional regularities; they can hide fine-grained semantic overlaps. The fact that the system uses pre-trained embeddings is another drawback because it can render it unable to capture the newly developed jargon in the industry unless periodically fine-tuned [15]. To overcome these limitations, it will be necessary to involve ongoing learning models and increase domain-specific corpora.

VI. CONCLUSION AND FUTURE WORK

This work proposes an adaptive hybrid clustering system that combines cross-language BERT embedding with a semantically oriented refinement layer to improve the classification of queries in multilingual business contexts. The experiment showed stable enhancements compared to baseline techniques, on Adjusted Rand Index (ARI), which improved between 0.74 and 0.86, and Silhouette Score, which improved between 0.65 and 0.78, representing a better purity and separation of clusters. The overall effect of the mixed use of the deep learning and statistical weighting methodology, regarding semantic ambiguity, was a significant reduction in the misclassification rates of domain-specific business vocabulary identified in the confusion matrix analysis. Scalability tests demonstrated that the system could handle large volumes of queries with comparatively low latency on the order of 220ms as opposed to 180ms per query at 100 queries per second, and is, therefore, suitable behind enterprise-grade applications requiring real-time capabilities. Overall, these findings fulfill the purpose of the research and justify the effectiveness of the hybrid method offered for multilingual customer service optimization. The resulting superior accuracy and scalability in the proposed system result in considerable improvements in operation within organizations with global clientele. In multilingual call centers and chatbot platforms, higher clustering efficiency is promised so that customers' questions will be sent to support venues that will help them in the best possible way, making their service process shorter and more satisfying. The modular design of its model allows it to upgrade in increments that may include model upgrades to more recent embedding schemes or more effective cluster-based techniques, without having to complete a full system-wide redeployment, which minimises maintenance fees and downtime.

AUTHOR CONTRIBUTIONS

Conceptualization – Yanmei Deng and Ooi Boon Keat; methodology – Yanmei Deng and Ooi Boon Keat; formal analysis – Yanmei Deng and Ooi Boon Keat; investigation – Yanmei Deng; resources – Yanmei Deng; writing-original

draft preparation – Yanmei Deng and Ooi Boon Keat; writing-review and editing – Yanmei Deng and Ooi Boon Keat. All authors have read and agreed to the published version of the manuscript.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

FUNDING

This work was supported by the horizontal research project titled “Business Data Analysis Optimization and Language Service Support”(ky2024272) in 2024 at Hubei Polytechnic University.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

- [1] C. Binder, C. Neureiter, and A. Lüder, “Towards a domain-specific approach enabling tool-supported modelbased systems engineering of complex industrial Internet-of-Things applications,” *Systems*, vol. 9, no. 2, pp. 21, 2021, doi: 10.3390/systems9020021.
- [2] X. Chen, “Research on the application of multilingual natural language processing technology in smart home systems,” *Journal of Computer, Signal, and System Research*, vol. 2, no. 5, pp. 8-14, 2025, doi: 10.71222/fgtq1p35.
- [3] P. Singhal, R. Walambe, S. Ramanna, and K. Kotecha, “Domain adaptation: Challenges, methods, datasets, and applications,” *IEEE Access*, vol. 11, pp. 6973-7020, 2023, doi: 10.1109/ACCESS.2023.3237025.
- [4] A. Petukhova, J. P. Matos-Carvalho, and N. Fachada, “Text clustering with large language model embeddings,” *International Journal of Cognitive Computing in Engineering*, vol. 6, pp. 100-108, 2025, doi: 10.1016/j.ijcce.2024.11.004.
- [5] S. M. Miraftabzadeh, C. G. Colombo, M. Longo, and F. Foiadelli, “K-means and alternative clustering methods in modern power systems,” *IEEE Access*, vol. 11, pp. 119596-119633, 2023, doi: 10.1109/ACCESS.2023.3327640.
- [6] S. Schmitz, U. Weidner, H. Hammer, and A. Thiele, “Evaluating uniform manifold approximation and projection for dimension reduction and visualization of PolInSAR features,” *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 1, pp. 39-46, 2021, doi: 10.5194/isprs-annals-V-1-2021-39-2021.
- [7] S. Wehrli, B. Arnrich, and C. Irrgang, “German text embedding clustering benchmark,” in Georges M, Herygers A, Friedrich A, Roth B, editors. *Proceedings of the 19th Conference on Natural Language Processing (KON-VENS 2023)*, 2023 Sep, Ingolstadt, Germany. Stroudsburg, PA, USA: Association for Computational Linguistics, 2023, pp. 187-201, doi: 10.48550/arXiv.2401.02709.
- [8] A. Mechergui, W. B. A. Karaa, and S. Zghal, “Cross-domain multilingual clustering: A generative hybrid model for constructing and enhancing semantic graphs from heterogeneous data,” *SN Computer Science*, vol. 5, no. 8, pp. 1066, 2024, doi: 10.1007/s42979-024-03374-3.
- [9] M. Umer, Z. Imtiaz, M. Ahmad, M. Nappi, C. Medaglia, G. S. Choi, et al., “Impact of convolutional neural network and FastText embedding on text classification,” *Multimedia Tools and Applications*, vol. 82, no. 4, pp. 5569-5585, 2023, doi: 10.1007/s11042-022-13459-x.
- [10] J. Myung, N. Lee, Y. Zhou, J. Jin, R. Putri, D. Antypas, et al., “Blend: A benchmark for LLMs on everyday knowledge in diverse cultures and languages,” *Advances in Neural Information Processing Systems*, 2024, pp. 78104-78146, [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2024/hash/8eb88844dafefa92a26aaec9f3acad93-Abstract-Datasets_and_Benchmarks_Track.html.
- [11] A. Fesseha, S. Xiong, E. D. Emiru, M. Diallo, and A. Dahou, “Text classification based on convolutional neural networks and word embedding for low-resource languages: Tigrinya,” *Information*, vol. 12, no. 2, pp. 52, 2021, doi: 10.3390/info12020052.
- [12] T. Sun, Z. Shu, Y. Huang, H. Wang, and Z. Yu, “Semantic feature graph consistency with contrastive cluster assignments for multilingual document clustering,” *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 24, no. 1, pp. 1-22, 2025, doi: 10.1145/3708887.
- [13] R. Piperno, L. Bacco, F. Dell’Orletta, M. Merone, and L. Pecchia, “Cross-lingual distillation for domain knowledge transfer with sentence transformers,” *Knowledge-Based Systems*, vol. 311, Art. no. 113079, 2025, doi: 10.1016/j.knosys.2025.113079.
- [14] Q. Zhang, Z. Yang, Y. Huang, Z. Chen, Z. Cai, K. Wang, et al., “A semantic alignment system for multilingual query-product retrieval,” *arXiv preprint arXiv:2208.02958*, 2022, doi: 10.48550/arXiv.2208.02958.
- [15] A. H. Mohammed and A. H. Ali, “Survey of BERT (bidirectional encoder representation transformer) types,” *Journal of Physics: Conference Series*, 2021. p. 012173, doi: 10.1088/1742-6596/1963/1/012173.